

**Measuring School and
Teacher Value Added in DC,
2011-2012 School Year**

Final Report

August 31, 2012

Eric Isenberg
Heinrich Hock



MATHEMATICA
Policy Research

Mathematica Reference Number:
06860.501

Submitted to:
Office of the State Superintendent of
Education
810 1st Street, NE
Washington, DC 20002
Project Officer: Robin Chait

District of Columbia Public Schools
1200 1st Street, NE
Washington, DC 20002
Project Officer: Olivia Meeks

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Eric Isenberg

Measuring School and Teacher Value Added in DC, 2011-2012 School Year

Final Report

August 31, 2012

Eric Isenberg
Heinrich Hock

MATHEMATICA
Policy Research

ACKNOWLEDGMENTS

We are grateful to the many people who contributed to this report. First, we would like to thank the Office of the State Superintendent of Education of the District of Columbia (OSSE) and the District of Columbia Public Schools (DCPS) for funding the work. We acknowledge Robin Chait and Jeffrey Noel at OSSE; Olivia Meeks, Cate Swinburn, and Jason Kamras at DCPS; and OSSE's Technical Support Committee, composed of representatives of six DC local education agencies (Joshua Boots, KIPP PCS; Anna Gregory, DCPS; Zachary Morford, Friendship PCS; Alexandra Pardo, Thurgood Marshall PCS; Richard Pohlman, E.L. Haynes PCS; and Megan Reamer, Capital City PCS) for working together to build a value-added model that meets the needs of the DC schools.

At Mathematica Policy Research, Mary Grider, assisted by Emma Kopa, Dejene Ayele, Juha Sohlberg, Jeremy Page, Dylan Ellis, and Becca Wang, processed the data and provided expert programming. Duncan Chaplin provided valuable comments. Carol Soble edited the report, and Jackie McGee and LaTia Downing provided word-processing and production support.

CONTENTS

I	OVERVIEW.....	1
	A. Using Value Added to Measure Performance	1
	B. A Value-Added Model for DC	2
	C. Caveats	5
II	DATA.....	7
	A. Teacher, School, and Student Lists	7
	B. DC CAS Test Scores	7
	C. Student Background Data	10
	D. School and Teacher Dosage.....	12
	1. School Dosage	13
	2. Teacher Dosage	13
III	ESTIMATING VALUE ADDED	15
	A. Regression Estimates.....	15
	1. The School Regression Model	15
	2. The Teacher Regression Model	16
	B. Measurement Error in the Pre-Tests	18
	C. Combining Estimates Across Grades	19
	D. Shrinkage Procedure.....	21
	E. Translating Value-Added Results to Scores for Evaluation Systems.....	22
	REFERENCES	23

TABLES

II.1	Reasons That Tested Students Were Excluded from the Analysis File: Math	8
II.2	Reasons That Tested Students Were Excluded from the Analysis File: Reading	9
II.3	Characteristics of Students in the Math Value-Added Models.....	11
II.4	Characteristics of Students in the Reading Value-Added Models.....	12
II.5	DCPS Schools Receiving Value-Added Estimates, by Grade Configuration	13
II.6	Teachers Receiving Value-Added Estimates, by Subject and Extent of Co-Teaching	14
III.1	Coefficients on Covariates in the Teacher and School Value-Added Models, by Subject.....	17
III.2	Student-Weighted Standard Deviations of Value-Added Estimates	20

I. OVERVIEW

In this report, we describe the value-added models used as part of teacher evaluation systems in the District of Columbia Public Schools (DCPS) and in eligible DC charter schools participating in Race to the Top. We estimated (1) teacher effectiveness in DCPS and eligible DC charter schools during the 2011–2012 school year and (2) school effectiveness in DCPS during the same year. This report updates our earlier technical reports (Isenberg and Hock 2010; Isenberg and Hock 2011).

The 2011–2012 school year is the third year of IMPACT, a teacher evaluation system for DCPS that relies in part on value-added estimates of teacher and school effectiveness. Under IMPACT, teachers who earn a highly effective rating have received performance pay, and those who earn an ineffective or a minimally effective rating for two consecutive years have been dismissed. Fifty percent of the total evaluation score for eligible teachers is composed of an individual teacher value-added score, and five percent of the total evaluation score of most teachers (including those ineligible to receive an individual value-added score) is composed of a school value-added score.

As part of a Race to the Top grant, 26 DC charter school local education agencies (LEA) have joined DCPS in using a teacher-level value-added model in their teacher evaluation systems for the 2011–2012 school year. Although each charter LEA has designed its own evaluation system for its personnel, value-added estimates will account for 50 percent of eligible charter school teachers' overall evaluation scores throughout participating LEAs. DC charter schools will use teacher evaluation scores to inform personnel decisions, which may include offering teachers performance pay for outstanding performance or dismissing teachers for poor performance.

A single value-added model that includes school, teacher, and student data from both DCPS and DC charter schools was used to generate estimates for all DC teachers. The DC Office of the State Superintendent of Education (OSSE) is managing the data on behalf of DC charter schools. OSSE, DCPS, and charter school LEAs agreed to make no substantive changes in how value added was calculated on the basis of data from the 2011–2012 school year compared to the method of calculation in the previous year.

In this report, we provide an overview of value-added methods in nontechnical terms (remainder of Chapter I); update the earlier technical reports by describing the data used to estimate teacher and school value added, including the new data on charter school students and teachers (Chapter II); and then describe the statistical methods used to estimate teacher and school value added (Chapter III). Chapters II and III include tables of diagnostic information that summarize the population of students and teachers on which the value-added estimates are based, as well as the results from the statistical model used to produce those estimates.

A. Using Value Added to Measure Performance

Value added is a method of measuring teacher effectiveness that seeks to isolate how much a teacher contributes to student achievement from any confounding factors outside the teacher's control.¹ To measure the performance of DC teachers, we use test scores and other data in a

¹ This section focuses on teacher value added, but the discussion applies equally to school value added. Along with teacher value added, we also estimated the value added of DCPS schools, which is incorporated into IMPACT. Subsequent footnotes highlight the differences between the teacher and school models.

statistical model designed to capture the achievement of students attributable to a teacher compared to the progress the students would have made with the average teacher. Known as a “value-added model” because it isolates the teacher’s contribution from other factors, this method has been developed and employed by a number of prominent researchers (Meyer 1997; Sanders 2000; McCaffrey et al. 2004; Raudenbush 2004; Hanushek et al. 2007) and is used to measure the performance of schools and/or teachers in many districts, including Chicago, Los Angeles, and New York City. Spurred in some cases by the federal government’s Race to the Top initiative, whole states have adopted value-added models to measure teacher performance, including Florida, New York, Pennsylvania, and Tennessee.

The basic approach of a value-added model is to predict the test scores that each student would have obtained with the average DC teacher and then compare the average actual scores of a given teacher’s students to the average predicted scores. The difference between these two scores—how the students actually performed with a teacher versus how they would have performed with the average DC teacher—represents the teacher’s “value added” to student achievement. For example, suppose that a 6th-grade math teacher has a class of students who, given their background characteristics such as poverty status, disability status, and test scores on the 5th-grade math and reading tests (or “pre-tests”), typically end the year 5 points above the citywide average on the 6th-grade math test (or “post-test”). The value-added model derives a relative measure of the teacher’s effectiveness by comparing the average student post-test score to the average predicted score. In this example, if the class post-test average is exactly 5 points above average, the value-added model will identify the teacher as an average performer. If the post-test average exceeds this standard, the teacher will be identified as above average, and if the average is less than the standard, the teacher will be considered below average. Because a value-added model accounts for students’ initial performance and other background characteristics, it allows any teachers to be identified as high performers, regardless of whether their students were low or high performing at baseline.

B. A Value-Added Model for DC

Although conceptually straightforward, the production of value-added estimates requires (1) the assembly of an analysis file of data from multiple sources and (2) the design of a value-added model that addresses several layers of complexity within DC’s educational context to accurately and fairly measure teachers’ performance. We briefly describe the key elements of the analysis file (described more fully in Chapter II) and then provide an overview of the steps used to estimate value added (see Chapter III for details).

We estimated the performance of teachers in DCPS and in participating charter schools using a value-added model based on the District of Columbia Comprehensive Assessment System (DC CAS) tests in math and reading. We measured teacher effectiveness in these two subjects separately. Similar to previous estimates we have calculated using only DCPS teachers, we calculated value added for teachers of math and/or English language arts/reading in grades 4–8 during the most recent school year, 2011–2012.

Elementary and middle school students were eligible to be included in the model if they had a post-test from 2012 and a pre-test from the same subject in the previous grade in 2011. We excluded grade repeaters so that the predicted achievement for all students in a grade was based on the same post-test and pre-test, allowing for meaningful comparisons between teachers. However, because of

this, not all students were included in the value-added model.² In addition to collecting post-test and pre-test scores, we collected data on student background characteristics.

When students change schools mid-year, teachers at multiple schools are responsible for their academic growth. To credit the teacher at one school with complete responsibility for a student who changes schools, or to ignore that student entirely, would distort our results. We therefore used school enrollment data to allocate proportional credit based on the fraction of time the student spent at each school. This fraction can be thought of as the “dosage” for the school.³

So that we could accurately match teachers to students, teachers eligible to receive value-added estimates participated in a process known as roster confirmation, in which they indicated whether and for how long they taught the students listed on their administrative rosters. We used the combination of school-enrollment information and roster-confirmed lists to create a dosage for each teacher-student pair, indicating whether a teacher taught a given student, and, if so, for how long.

After constructing the analysis file, we estimated the value-added model using four steps, each of which addressed a different conceptual challenge.

1. **Multiple regression.** We used multiple regression, a statistical technique that allowed us to simultaneously account for a group of background factors to avoid holding teachers accountable for factors outside their control. We accounted for a set of student characteristics that could be related to performance on the DC CAS test in the 2011–2012 school year (the post-test). These characteristics include a student’s DC CAS test scores from the 2010–2011 school year in math and reading (the pre-tests), poverty status, limited English proficiency, special education status, and school attendance during the 2010–2011 school year.^{4,5} Accounting for these characteristics, we obtained a net estimate of each teacher’s effectiveness. The estimate is approximately equal to the difference between the average actual post-test score of a teacher’s students and the average predicted score of those students based on their characteristics. We weighted each student’s contribution to a teacher’s score by the proportion of time the student was assigned to the teacher when the teacher was teaching that subject.

² Only DCPS students in grades 4–8 and 10 are eligible for inclusion in the school value-added model. Most 10th-grade students took the 8th-grade DC CAS tests in spring 2010, but about three percent took these tests in spring 2009. We therefore used tests with a two- or three-year lag between pre-test and post-test for 10th-grade students.

³ For the school value-added model, we measured the school dosage for 10th-grade students over two years because most students took the pre-test two years earlier.

⁴ In the school model, we do not account for attendance in the prior year but, in case there might have been a systematic difference in achievement between 10th-grade students who took the 8th-grade test in 2009 or 2010, we account for whether students took an extra year to progress from 8th- to 10th-grade. We estimate a different relationship between pre-test and post-test scores for these two groups of students.

⁵ A student’s race/ethnicity or gender may be correlated with factors that both affect test scores and are beyond a teacher’s control. Preliminary results showed a high correlation between value-added measures estimated with and without race/ethnicity or gender. This suggests that the other characteristics included in the value-added models capture most of the factors affecting test scores that are correlated with race/ethnicity and gender. As a result, DCPS decided not to account for race/ethnicity and removed gender from the value-added model after the first year of IMPACT.

We developed a technique known as the “Full Roster Method” to estimate the combined effectiveness of teachers when they shared students (Hock and Isenberg 2012).⁶ In some cases, a student was taught by one teacher for part of the year and another teacher for the remainder of the year. In other cases, two or more teachers were jointly responsible for a classroom of students at the same time. Teachers of shared students received equal credit for their achievement.

2. **Accounting for measurement error in the pre-test.** Given that a student’s performance on a single test is an imperfect measure of ability, teachers may be unfairly held accountable for the initial performance of their students, rather than being assessed on the gains they produce in student learning. Good or bad luck on the pre-test can dampen the observed relationship between pre- and post-test scores, compared to the true relationship between student achievement at the beginning and end of the year. So, if we were to use the observed relationships without any adjustments, teachers of students with low pre-test scores might be held partly accountable for the performance of their students before they entered their classrooms. To avoid this problem, we compensated for good or bad luck in pre-test scores—also known as measurement error—by employing a statistical technique that uses published information on the test/retest reliability of the DC CAS.
3. **Comparing teachers across grades.** The DC CAS is not designed to allow the comparison of scores across grades. We therefore placed teachers on a common scale by translating each teacher’s value-added estimate into a metric of “generalized” DC CAS points. This translation was based on a three-stage procedure. First, before the multiple-regression step, we translated student math and reading scores into a common metric in which each student test score is measured relative to other test scores within the same year, grade, and subject. These scores were then used to produce initial teacher value-added estimates. Second, we adjusted these estimates so that the average teacher in each grade receives the same score. Third, we multiplied the resulting estimates by a grade-specific conversion factor to ensure that the dispersion of teacher value-added estimates is similar by grade. For teachers with students in more than one grade, we took a student-weighted average of their grade-specific value-added estimates.⁷
4. **Accounting for imprecisely estimated measures based on few students.** Estimates of teacher effectiveness could be misleading if based on too few students. Some students may score high on a test due to good luck rather than good preparation, and others may score low due to bad luck. For teachers with many students, good and bad luck affecting test performance tends to cancel out. A teacher with few students, however, can receive a very high or very low effectiveness rating based primarily on luck (Kane and Staiger 2002). We reduced the possibility of such spurious results by (1) not reporting estimates for teachers with fewer than 15 students⁸ and (2) using a statistical technique that

⁶ In the school value-added model, we used a technique known as the Partial Credit Method to estimate value added when students are taught by more than one school. This method attributes student achievement differently across schools for the same student (Hock and Isenberg 2012).

⁷ To compare schools with different grade configurations, we applied a similar strategy. We transformed each grade-level estimate within a school into generalized DC CAS points and then averaged the grade-level estimates across grades to arrive at a composite value-added estimate for the school.

⁸ For schools, the threshold is 100 test scores across math and reading. For example, a school would qualify if it had 50 eligible students, all of whom took both the math and reading test.

combines the effectiveness measure of a particular teacher (from step 3) with the overall average to produce a final value-added estimate (Morris 1983). We relied more heavily on a default assumption of average effectiveness for teachers with few students or with students whose achievement is most difficult to predict with a statistical model.

C. Caveats

It is important to recognize the limitations of any performance measures, including those generated by a value-added model. Below, we discuss three caveats that are especially important for interpreting and using the results of a value-added model like the one we created for DC.

1. **Estimation error.** The value-added measures are estimates of a teacher's performance based on the available data and the value-added model used. As with any statistical model, there is uncertainty in the estimates produced, which implies that two teachers with similar value-added estimates are “statistically indistinguishable” from one another. We quantified the precision with which the measures were estimated by reporting the upper and lower bounds of a confidence interval of performance for each teacher.
2. **Unmeasured differences between students.** A value-added model uses statistical techniques to account (or “control”) for differences in student performance based on documented sources of information about students, such as their prior-year test score or free-lunch eligibility. However, the model cannot control for differences in student performance that arise from sources that are not explicitly measured. For this reason, policymakers may have concerns about how to interpret value-added estimates. For example, one concern might be that teachers at certain schools would be unfairly rewarded if especially motivated parents choose schools for their children in ways that are not accounted for by the student characteristics the value-added model. Similarly, if the assignment of students to teachers within schools was based on unobservable factors—for example, pairing difficult-to-teach students with teachers who have succeeded with similar students in the past—a value-added model might unfairly penalize these teachers because it cannot statistically account for such factors. A related concern is that teacher-level value added might reflect the efficacy of school inputs, such as the leadership of the principal or a consistent, schoolwide student behavior code. Partly for these reasons, some researchers caution against using value-added models (Rothstein 2010), especially for comparing teachers across schools (Aaronson et al. 2007).

Empirical work in experimental settings (Kane and Staiger 2008) and quasi-experimental settings (Chetty et al. 2011) suggests these factors do not play a large role in determining teacher value added. Using data from the Los Angeles Unified School District, Kane and Staiger (2008) compared (1) the difference in value-added measures between pairs of teachers based on a typical situation in which principals assign students to teachers and (2) the difference in student achievement between the teachers the following year, when they taught classrooms that were formed by principals but then randomly assigned to the teachers. Kane and Staiger found that the differences between teachers' value-added estimates before random assignment were a statistically significant predictor of achievement differences when classrooms were assigned randomly. Because these results were gathered in schools in which the principal was willing to allow random assignment of classrooms to teachers, it is not clear whether they would generalize to other settings. Chetty et al. (2011), however, complemented these findings, using longitudinal data from a large urban district. They showed that the value added of teachers who change schools

persists in their new settings. This suggests that value added reflects a teacher's performance in the classroom, not some unmeasured characteristic of the teacher's students.

3. **Single versus multiple measures.** Value-added estimates measure a teacher's contribution to student achievement based on standardized test scores. Additional measures of teacher effectiveness may improve the predictive power of teacher evaluation systems (Kane and Staiger 2012) or the future effectiveness of teachers (Taylor and Tyler 2011). Although evaluation systems vary from LEA to LEA, all LEAs participating in Race to the Top in DC use multiple measures of teacher effectiveness. In addition to value added, some include components designed to capture effective lesson planning and instructional delivery; a teacher's commitment to the school community; and the fostering of a positive environment for students.

II. DATA

In this chapter, we review the data used to generate the value-added measures. We discuss the data on student background characteristics and how we calculate the amount of time that students spent in more than one school or with more than one teacher. We provide an overview of the roster confirmation process that allows teachers to confirm whether and for what portion of the year they taught math and/or reading to students. Finally, we discuss the standardized assessment used in DC schools.

A. Teacher, School, and Student Lists

OSSE provided an official comprehensive list of charter schools with eligible teachers. The schools participate in Race to the Top and include at least some of the grades from 4 to 8. DCPS provided an official comprehensive list of schools eligible for inclusion in the school value-added model as well as a list of teachers of math and reading in grades 4 through 8 who are eligible to receive individual value-added estimates. In general, only regular education teachers are eligible to receive value-added estimates; special education teachers are not eligible to receive value-added estimates.⁹ OSSE provided the data from which to construct a student list that indicates charter school students' official grade level. DCPS provided a list of official grade levels for DCPS students. For the school model, we used only data for DCPS students. For the teacher model, we used data for DCPS and charter school students.

B. DC CAS Test Scores

When estimating the effectiveness of DC teachers, we included elementary and middle school students if they have a DC CAS test from 2012 (the post-test) and a DC CAS test from the previous grade in the same subject in 2011 (the pre-test).¹⁰ We excluded students from the analysis file in the case of missing or conflicting school enrollment data, test-score data, or student background data.^{11,12} We also excluded students who repeated or skipped a grade because they lack pre-test and post-test scores in consecutive grades and years. When estimating the effectiveness of DCPS schools, we included all DCPS students in grades 4 to 8 who are in the teacher analysis file, plus DCPS students in grade 10 if they have a pre-test from grade 8 in the same subject in either 2009 or 2010.

Table II.1 shows how the analysis file was constructed for teacher and school value-added models in math; Table II.2 shows the same information for reading. The first two columns in Tables II.1 and II.2 show the totals for the teacher models, using data for DCPS and charter school students in grades 4–8, and the last two columns show the totals for the school models, using only data for DCPS students for grades 4–8 and 10. Tables II.1 and II.2 show the total number of students who could have been included in the analyses, the reasons why students were excluded, and

⁹ DCPS also included some resource and Read 180 teachers among the eligible teachers.

¹⁰ DCPS provided us with DC CAS test scores for DCPS students from 2009 to 2011. OSSE provided us with DC CAS scores for charter school students from 2011, and for all DC students in 2012.

¹¹ We considered a student who answered fewer than five questions on the DC CAS post-test to be missing test score data.

¹² We included students who are missing individual student background characteristics but excluded those for whom no data on background characteristics were available.

the total included in the models. The top row of these tables shows the total number of students who received test scores for the 2012 DC CAS test. The next four rows show the reasons why students who had 2012 test scores were excluded from the analysis file. As shown in the bottom rows of Tables II.1 and II.2, over 90 percent of students with 2012 test scores were included in the teacher value-added models.

Table II.1. Reasons That Tested Students Were Excluded from the Analysis File: Math

	Teacher Value-Added Model: DCPS and Charter Schools		School Value-Added Model: DCPS Only	
	Number	Percent	Number	Percent
Students with Math Post-test Scores in 2012	21,086	100.0	16,231	100.0
(1) Student has conflicting school enrollment	31	0.1	3	0.0
(2) Tested grade level for math differed from official grade level	28	0.1	16	0.1
(3) Missing math pre-test scores	1,545	7.3	1,707	10.5
(4) Skipped or repeated a grade	364	1.7	198	1.2
Total Excluded	1,968	9.3	1,924	11.9
Total Included in Math Value-added Models	19,118	90.7	14,307	88.1

Notes: Students are not included in this table if they only enrolled in charter schools not participating in Race to the Top or if they cannot be linked to background characteristics.

Students are excluded sequentially in the order presented and so do not count for more than one reason in this table.

A student's tested grade level is based on the first digit of the DC CAS post-test, whereas the official grade level is drawn from administrative records at DCPS and the Public Charter School Board.

The teacher value-added model includes DCPS and charter school students in grades 4–8. The school value-added model includes DCPS students in grade 10 and excludes charter school students in grades 4–8. The school model therefore includes DCPS students in grades 4–8 and grade 10.

Unlike students in 4th through 8th grade, eligibility for inclusion as a 10th-grade student is based on having an 8th-grade test score in the 2009 or 2010 school years as well as a 10th-grade test score in 2012. Students with an 8th-grade test score from any other school year are counted as having missing pre-test scores.

Table II.2. Reasons That Tested Students Were Excluded from the Analysis File: Reading

	Teacher Value-Added Model: DCPS and Charter Schools		School Value-Added Model: DCPS Only	
	Number	Percent	Number	Percent
Students with Reading Post-Test Scores in 2012	21,005	100.0	16,219	100.0
(1) Student has conflicting school enrollment	31	0.1	3	0.0
(2) Tested grade level for reading differed from official grade level	31	0.1	55	0.3
(3) Missing reading pre-test scores	1,614	7.7	1,758	10.8
(4) Skipped or repeated a grade	366	1.7	201	1.2
Total Excluded	2,042	9.7	2,017	12.4
Total Included in Reading Value-Added Models	18,963	90.3	14,202	87.6

Notes: Students are not included in this table if they only enrolled in charter schools not participating in Race to the Top or if they cannot be linked to background characteristics.

Students are excluded sequentially in the order presented and so do not count for more than one reason in this table.

A student's tested grade level is based on the first digit of the DC CAS post-test, whereas the official grade level is drawn from administrative records at DCPS and the Public Charter School Board.

The teacher value-added model includes DCPS and charter school students in grades 4–8. The school value-added model includes DCPS students in grade 10 and excludes charter school students in grades 4–8. The school model therefore includes DCPS students in grades 4–8 and grade 10.

Unlike students in 4th through 8th grade, eligibility for inclusion as a 10th-grade student is based on having an 8th-grade test score in the 2009 or 2010 school years as well as a 10th-grade test score in 2012. Students with an 8th-grade test score from any other school year are counted as having missing pre-test scores.

To obtain precise estimates of teacher effectiveness, we estimated the value-added model for teachers by using all students in grades 4 through 8, even those not linked to an eligible teacher, because they (1) were not taught by a DC teacher for at least 5 percent of the school year, (2) were included in the roster file but not claimed by a teacher, or (3) were claimed only by a teacher with fewer than 7 students in his or her grade (we do not estimate a value-added measure for teachers with so few students). Inclusion of these unlinked students in the analysis allowed us to estimate more precisely the relationship between student characteristics and achievement for all students, including the majority of students linked to teachers. We report estimates only for teachers who taught 15 or more students over the course of the year in at least one subject.

For each subject, the DC CAS is scored so that each student receives a scale score from 300 to 399 for grade 3 students, 400 to 499 for grade 4 students, and so on. The range for grade 10 students is 900 to 999. The first digit is a grade indicator only and does not reflect student achievement. We dropped the first digit and used the rest of the score, which ranges from 0 to 99.

The resulting scores may be meaningfully compared only within grades and within subjects; math scores, for example, are generally more dispersed than reading scores within the same grade. Therefore, before using the test scores in the value-added model, we created subject- and grade-

specific z-scores by subtracting the mean and dividing by the standard deviation within a subject-grade combination.¹³ This step allowed us to translate math and reading scores in every grade and subject into a common metric. To create a measure with a range resembling the original DC CAS-point metric, we then multiplied each test score by the average standard deviation across all grades within each subject and year.

C. Student Background Data

We used data provided by OSSE and DCPS to construct variables used in the value-added models as controls for student background characteristics. In both the school and teacher value-added models, we controlled for:

- Pre-test in same subject as post-test
- Pre-test in other subject (we control for math and reading pre-tests regardless of post-test)
- Free-lunch eligibility
- Reduced-price lunch eligibility
- Limited English proficiency status
- Existence of a specific learning disability
- Existence of other types of disabilities requiring special education

In the teacher model, we also controlled for:

- Proportion of days that the student attended school during the previous year

In the school model, we also controlled for:

- Taking the grade 8 DC Comprehensive Assessment System (CAS) test in 2009 rather than in 2010 (for some grade 10 students)

Attendance is a measure of student motivation. We used previous—rather than current-year—attendance to avoid confounding student attendance with current-year teacher effectiveness; that is, a good teacher versus a weaker teacher might be expected to motivate students to attend school more regularly. We excluded the attendance measure from the school model because many students will have attended the same school in the previous year, thereby posing a risk for confounding student attendance with school effectiveness. The last variable in the list is relevant only for the school model because we did not include 10th-grade students in the teacher model. Attendance is a continuous variable that could range from zero to one. Aside from pre-test variables, the other variables are binary variables taking the value zero or one. The tables below show the characteristics of students included in the math value-added models (Table II.3) and reading value-added models (Table II.4).

¹³ Subtracting the mean score for each subject and grade creates a score with a mean of zero in all subject-grade combinations.

Table II.3. Characteristics of Students in the Math Value-Added Models

	Teacher Value-Added Model: DCPS and Charter Schools		School Value-Added Model: DCPS Only	
	Number	Percent	Number	Percent
Included in math value-added models	19,118	100.0	14,307	100.0
Linked to an eligible school for any part of the year	18,849	98.6	13,948	97.5
Linked to a teacher in the math value-added model for any part of the year	17,476	91.4	n.a.	n.a.
Eligible for free lunch	13,252	69.3	9,632	67.3
Eligible for reduced-price lunch	1,184	6.2	714	5.0
Limited English proficiency	1,260	6.6	905	6.3
Specific learning disability	1,755	9.2	1,423	9.9
Other learning disability	1,633	8.5	1,192	8.3
Took additional time between grades 8 and 10	n.a.	n.a.	394	2.8

Notes: The counts in the top row correspond to the final row of Table II.1.

All percentages are based on the counts in the top row.

Student characteristics were calculated as a weighted average for students enrolled in both a DCPS and charter school. The counts and percentages were not weighted in any other way.

Teachers in DCPS and charter schools are eligible for the teacher value-added model, but only DCPS schools are eligible for the school value-added model.

Free and reduced-price lunch status was imputed using data from prior years for approximately 12 percent of students in the teacher model and 18 percent of students in the school model. For all other student characteristics, less than 1 percent of students have missing data.

Students who took additional time between grades 8 and 10 are 10th-grade students in the 2011–2012 school year who took three years to progress from grade 8 to grade 10.

We imputed data for students who were included in the analysis file but who had missing values for one or more student characteristics. Our imputation approach used the values of nonmissing student characteristics to predict the value of the missing characteristic.¹⁴ For students who did not attend a DC school for part of the previous year, we used a Bayesian method to impute missing attendance data based on other student characteristics in addition to

¹⁴ For missing data on free or reduced-price lunch status, we used a slightly more sophisticated imputation procedure because these data are missing for DCPS students attending Provision 2 schools, which do not collect information on free and reduced-price lunch status every year. For these students, we used direct certification data to determine free-lunch status. If a student in a Provision 2 school is not eligible to receive a free lunch via direct certification, we used the student's status from a prior year. If no prior-year data are available for these students, we predicted the value of free and reduced-price lunch status using (1) the percentage of students eligible to receive a free or reduced-price lunch in the school in the last year during which the school sought to collect this information from all students and (2) individual student characteristics that are correlated with free-lunch status within schools.

attendance during the portion of the year spent in DC.¹⁵ We did not generate imputed values for the same-subject pre-test; we dropped from the analysis file any students with missing same-subject pre-test scores.

Table II.4. Characteristics of Students in the Reading Value-Added Models

	Teacher Value-Added Model: DCPS and Charter Schools		School Value-Added Model: DCPS Only	
	Number	Percent	Number	Percent
Included in reading value-added models	18,963	100.0	14,202	100.0
Linked to an eligible school for any part of the year	18,690	98.6	13,838	97.4
Linked to a teacher in the reading value-added model for any part of the year	17,163	90.5	n.a.	n.a.
Eligible for free lunch	13,155	69.4	9,562	67.3
Eligible for reduced-price lunch	1,175	6.2	704	5.0
Limited English proficiency	1,142	6.0	794	5.6
Specific learning disability	1,753	9.2	1,429	10.1
Other learning disability	1,635	8.6	1,202	8.5
Took additional time between grades 8 and 10	n.a.	n.a.	400	2.8

Notes: The counts in the top row correspond to the final row of Table II.1.

All percentages are based on the counts in the top row.

Student characteristics were calculated as a weighted average for students enrolled in both a DCPS and charter school. The counts and percentages were not weighted in any other way.

Teachers in DCPS and charter schools are eligible for the teacher value-added model, but only DCPS schools are eligible for the school value-added model.

Free and reduced-price lunch status was imputed using data from prior years for approximately 12 percent of students in the teacher model and 18 percent of students in the school model. For all other student characteristics, less than 1 percent of students have missing data.

Students who took additional time between grades 8 and 10 are 10th-grade students in the 2011–2012 school year who took three years to progress from grade 8 to grade 10.

D. School and Teacher Dosage

Given that some students moved between schools or were taught by a combination of teachers, we apportioned their achievement among more than one school or teacher. We refer to the fraction of time the student was enrolled at each school and with each teacher as the “dosage.”

¹⁵ We generated a predicted value by using the values of nonmissing student characteristics, and we combined this information with the actual attendance data for the part of the year spent in DC. With this method, the more time a student spent in a DC school, the more his or her imputed attendance measure relies on actual attendance data from the part of the year spent in DC. Conversely, the less time spent in DC, the more the imputed attendance measure relies on the predicted value. We implemented this approach by using a beta distribution with beta/binomial updating (Lee 1997).

1. School Dosage

Based on administrative data from OSSE and DCPS, which contain dates of school withdrawal and admission, we assigned every student a dosage for each school attended. School dosage equals the fraction of the first three quarters of the school year that a student was officially enrolled in that school. In determining dosage, we used school calendars from each participating LEA and only the first three quarters of the year because students in most LEAs start taking their tests shortly after the end of the third quarter.¹⁶ To account fully for each student's time, we also recorded the portion of the school year that the student was enrolled in schools outside DC.

Recognizing that a school is unlikely to have an appreciable educational impact on a short-term student, we set dosage equal to zero for students who spent less than 5 percent of the year at a school. Conversely, we set dosage to 100 percent for students who spent 95 percent or more of the year at a school. Apart from this, in the school model, we assumed that learning accumulated at a constant rate and therefore treated days spent at one school as interchangeable with days spent at another. For example, if a student split time equally between two schools, we set the dosage of each school to 50 percent regardless of which school the student first attended. Given that the 8th-grade DC CAS test is the pre-test for students in 10th grade, we based dosage variables for 10th-grade students on (1) the schools they attended during the 2010–2011 and 2011–2012 school years for students who took the 8th-grade DC CAS test in 2009 or on (2) the schools they attended during the 2009–2010 through 2011–2012 school years for students who took the 8th-grade DC CAS test in 2008. Table II.5 shows the number of schools receiving reported value-added estimates, by grade configuration.

Table II.5. DCPS Schools Receiving Value-Added Estimates, by Grade Configuration

	Number	Percent
Elementary schools	47	49.0
Middle schools	14	14.6
High schools	16	16.7
Education campuses	18	18.8
Other	1	1.0
Total	96	100.0

Notes: Schools received estimates if they are linked to at least 100 eligible students across the two subjects (for example, 50 students taking both math and reading tests).

2. Teacher Dosage

To determine which students received math and reading instruction from eligible teachers in 2011–2012, DC schools conducted a roster confirmation among teachers of math and reading in grades 4–8. In most cases, teachers received lists of students who appeared on their course rosters. Teachers could also add students to their rosters. In a few cases, including all teachers in one LEA,

¹⁶ There are two exceptions. One LEA uses trimesters rather than quarters. In a school at another LEA, the third quarter ends after the DC CAS tests are administered. In both of these cases, we used enrollment data from the first two terms and from the third term before the beginning of the testing window for the DC CAS.

teachers added all of their students because a prefilled roster was unavailable. For each of the first three quarters, teachers indicated whether they taught each subject to each student and, if so, the proportion of time they taught the student. For example, if a student spends two days a week in an eligible teacher's classroom learning math and three days per week in another classroom with a special education teacher while other students learn math with the eligible teacher, the student was recorded as having spent 40 percent of instructional time with the eligible teacher. In recording the proportion of time spent with a student in a given class and subject, teachers rounded to the nearest 20 percent such that the possible responses were 0, 20, 40, 60, 80, and 100 percent. If a teacher claimed a student for less than 100 percent in any quarter, the teacher was not responsible for naming other teachers who taught the student, but they did indicate the reason for the reduction. OSSE ensured that all eligible charter school teachers completed roster confirmation. Within each charter LEA, a central office administrator, principal, or designee verified the confirmed rosters. Likewise, in DCPS, principals verified confirmed rosters. Central office staff at DCPS also followed up with DCPS teachers as necessary.

We used the confirmed class rosters to construct teacher-student links. If the roster confirmation data indicated that a student had one math or reading teacher at a school, we set the teacher-student weight equal to the school dosage. If a student changed teachers from one term to another, we determined the number of days the student spent with each teacher, subdividing the school dosage among teachers accordingly. When two or more teachers claimed the same students during the same term, we assigned each teacher full credit for the shared students. This reflects a decision by OSSE and DCPS that solo-taught and co-taught students contribute equally to teachers' value-added estimates. We therefore did not subdivide dosage for co-taught students. Finally, similar to tracking time spent at all schools outside DC, we tracked the time a student spent with any teachers not recorded in the confirmed class rosters. Table II.6 shows how many teachers shared students with another teacher in the value-added model and what percentage of their students were shared. As shown in the table, 77.2 percent of math teachers and 65.1 percent of reading teachers shared none their students with another teacher, whereas 8.1 percent of math teachers and 11.9 percent of reading teachers shared all of their students with other teachers.

Table II.6. Teachers Receiving Value-Added Estimates, by Subject and Extent of Co-Teaching

Fraction of Students Shared with Another Teacher in Value-Added Model	Math		Reading	
	Number	Percent	Number	Percent
None	370	77.2	328	65.1
1-10 Percent	42	8.8	47	9.3
11-50 Percent	11	2.3	48	9.5
51-99 Percent	17	3.5	21	4.2
All Students	39	8.1	60	11.9
Total	479	100.0	504	100.0

Notes: Teachers received estimates if they were linked to at least 15 eligible students.

A co-teacher is any teacher who shares at least one student with another teacher in the value-added model.

III. ESTIMATING VALUE ADDED

A. Regression Estimates

We have developed two linear regression models to estimate effectiveness measures for schools and for teachers. After assembling the analysis file, we estimated a regression separately for math and reading using students at all grade levels in the data. In each regression equation, the post-test score depends on prior achievement, student background characteristics, variables linking students to schools or teachers, and unmeasured factors.

1. The School Regression Model

The regression equation used to estimate effectiveness measures for schools may be expressed formally as:

$$(1) \quad Y_{ig} = \lambda_{1g} Y_{i(g-1)} + \omega_{1g} Z_{i(g-1)} + \alpha'_1 \mathbf{X}_{1i} + \beta' \mathbf{S}_{ig} + \varepsilon_{1ig},$$

where Y_{ig} is the post-test score for student i in grade g and $Y_{i(g-1)}$ is the same-subject pre-test for student i in grade $g-1$ during the previous year. The variable $Z_{i(g-1)}$ denotes the pre-test in the opposite subject. Thus, when estimating school effectiveness in math, Y represents math tests with Z representing reading tests and vice versa. The pre-test scores capture prior inputs into student achievement, and the associated coefficients, λ_{1g} and ω_{1g} , vary by grade. The vector \mathbf{X}_{1i} denotes the control variables for individual student background characteristics. The coefficients on these characteristics, α_1 , are constrained to be the same across all grades.¹⁷

The vector \mathbf{S}_{ig} contains one dosage variable for each school-grade combination, and the associated coefficients contained in β measure the effectiveness of each school by grade. For students attending more than one school, each school received partial credit based on the student's dosage. The dosage for a given element of \mathbf{S}_{ig} is set to be equal to the percentage of the year student i was enrolled in grade g at that school. The vector of dosage variables (\mathbf{S}_{ig}) also includes a separate variable for each grade level for the fraction of the school year a student spent outside DCPS. The value of any element of \mathbf{S}_{ig} is zero if student i was not taught in grade g in that school during the school year. Because \mathbf{S}_{ig} accounts for student attendance throughout the school year, its elements always sum to one. Rather than dropping one of the school dosage variables from the regression, we estimated the model without a constant term. We also mean centered the control variables so that each element of β represents a school- and grade-specific intercept term for a student with average characteristics.¹⁸ We assumed that there are systematic differences in the variability of test outcomes

¹⁷ We estimated a common, grade-invariant set of coefficients of student background characteristics because our calculations using 2009–2010 data revealed substantial differences in sign and magnitude of grade-specific coefficients on these covariates. These cross-grade differences appeared to reflect small within-grade samples of individuals with certain characteristics rather than true differences in the association between student characteristics and achievement growth. Estimating a common set of coefficients across grades allowed us to base the association between achievement and student characteristics on information from all grades, which should smooth out the between-grade differences in these coefficients.

¹⁸ Mean centering the student characteristics and pre-test scores tends to reduce the estimated standard errors of the school effects (Wooldridge 2008).

across different types of students or at different schools, which implies that the error term, ε_{ig} , is heteroskedastic.

We estimated equation (1) by using ordinary least squares (OLS). Heteroskedasticity generally results in estimated errors that are too small because the regression does not account for all sources of variability. Accordingly, we calculated heteroskedasticity-robust standard errors using the Huber-White estimator (Huber 1967; White 1980).

The regression produced separate value-added coefficients for each grade within a school. To reduce the likelihood of obtaining statistically imprecise estimates, we did not include dosage variables for school-grade combinations with fewer than five student equivalents.¹⁹ We aggregated the estimated coefficients into a single measure for each school (see Section C below).

2. The Teacher Regression Model

The teacher model differs from the school model in order to account for team teaching that occurs at the teacher level but not at the school level. The chief difference in the teacher model is that the unit of observation is a teacher-student combination rather than a student. Unlike the school model, in which schools contribute separately to the achievement of students who attend more than one school, the teacher model is based on the assumption that the combined effort of team teachers constitutes a single input into student achievement (Hock and Isenberg 2012). For a given teacher t and student i , the regression equation may be expressed as:

$$(2) \quad Y_{tig} = \lambda_{2g} Y_{i(g-1)} + \omega_{2g} Z_{i(g-1)} + \alpha'_2 \mathbf{X}_{2i} + \eta' \mathbf{T}_{tig} + \varepsilon_{2tig},$$

where the notation largely parallels that for the school model described by equation (1). The vector \mathbf{T}_{tig} includes a grade-specific variable for each teacher and includes a variable for a catchall ineligible teacher in each grade to account for student dosage that cannot be attributed to a particular teacher who is eligible to receive a value-added estimate. A student contributes one observation to the model for each teacher to whom the student is linked, based on the roster confirmation process. Each teacher-student observation has one nonzero element in \mathbf{T}_{tig} . Measures of teacher effectiveness are contained in the coefficient vector η . Table III.1 shows the coefficients and standard errors of the control variables in the school and teacher models. The top panel of Table III.1 shows the association between a given student characteristic and the change in achievement on the 2011 DC CAS (measured in points on the test). The bottom panel shows the average association between the pre-tests and the post-test.

¹⁹ In practice, this is likely to occur only in auxiliary models in which we restrict the sample to students who belong to specific categories, such as special education. For school-grade combinations that do not meet the five-student-equivalent threshold, we will reassigned the dosage for these students to a variable representing time spent at all other schools in the given grade, including schools outside DCPS.

Table III.1. Coefficients on Covariates in the Teacher and School Value-Added Models, by Subject

Variable	Teacher Value-Added Model: DCPS and Charter Schools		School Value-Added Model: DCPS Only	
	Math	Reading	Math	Reading
Eligible for free lunch	-0.804 (0.196)	-1.378 (0.183)	-1.173 (0.234)	-1.259 (0.212)
Eligible for reduced-price lunch	-0.172 (0.292)	-0.898 (0.304)	-0.179 (0.380)	-0.193 (0.369)
Limited English proficiency	0.217 (0.324)	-0.665 (0.322)	0.994 (0.404)	-0.346 (0.398)
Specific learning disability	-2.067 (0.319)	-2.620 (0.297)	-2.154 (0.376)	-3.044 (0.349)
Other learning disability	-2.687 (0.345)	-3.336 (0.330)	-2.751 (0.425)	-3.463 (0.412)
Took additional time between grades 8 and 10	n.a.	n.a.	-13.757 (0.781)	-12.716 (0.744)
Fraction of the prior year student attended school	7.300 (1.540)	0.226 (1.361)	n.a.	n.a.
Pre-Test Scores (average coefficients)				
Same subject, all grades, standard time between grades	0.712 (0.020)	0.601 (0.019)	0.681 (0.026)	0.606 (0.028)
Opposite subject, all grades, standard time between grades	0.100 (0.024)	0.159 (0.017)	0.172 (0.031)	0.167 (0.024)
Same subject, grade 10, additional time between grades	n.a.	n.a.	0.337 (0.045)	0.502 (0.045)
Opposite subject, grade 10, additional time between grades	n.a.	n.a.	0.164 (0.050)	0.120 (0.040)

Notes: Standard errors are in parentheses.

The reported coefficient estimates of pre-test scores for all grades represent averages of the coefficients estimated separately for grades 4–8 in the teacher model and grades 4–8 and 10 in the school model. The associated standard errors similarly represent averages across grades. The standard errors therefore do not account for the variability of the estimates across grades. These numbers are presented for descriptive purposes only and should not be used to conduct statistical inference.

For students in grades 4–8, pre-test scores are from the prior grade in the 2010–2011 school year. For 10th-grade students, pre-test scores are (1) from 8th-grade test scores in the 2009–2010 school year for students who take two years to progress from 8th to 10th grade and (2) from 8th-grade test scores in the 2008–2009 school year for students who take three years to progress from 8th to 10th grade.

To account for multiple observations on the same student, we estimated the coefficients by using weighted least squares (WLS) rather than OLS. In this method, the teacher-grade variables in \mathbf{T}_{tig} are binary, and we weighted each teacher-student combination by the teacher dosage associated with that combination. We addressed the correlation in the error term, ε_{2tig} , across multiple observations by using a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano

1987) to obtain standard errors that are consistent in the presence of both heteroskedasticity and clustering at the student level.

Similar to the school model, the teacher regression yields separate value-added coefficients for each grade in which a teacher is linked to students. To improve the precision of the estimates, we estimated a grade-specific coefficient for a teacher only if he or she teaches at least seven students in that grade.²⁰ We then aggregated teacher estimates across grades to form a single estimate for each teacher (see Section C below).

B. Measurement Error in the Pre-Tests

We corrected for measurement error in the pre-tests by using grade-specific reliability data available from the test publisher (CTB/McGraw Hill 2009; 2010; 2011). As a measure of true student ability, standardized tests contain measurement error, causing standard regression techniques to produce biased estimates of teacher or school effectiveness. To address this issue, we implemented a measurement error correction based on the test/retest reliability of the DC CAS tests. By netting out the known amount of measurement error, the errors-in-variables correction eliminates this source of bias (Buonaccorsi 2010).

Correcting for measurement error requires a two-step procedure. Our statistical model includes distinct pre-test coefficients for each grade but common coefficients on student characteristics. However, it is not computationally possible to apply the numerical formula for the errors-in-variables correction simultaneously to all grades. Therefore, we estimated the errors-in-variables correction in the first step on a grade-by-grade basis and then estimated a second-step regression with common (rather than grade-specific) coefficients on the student characteristics. We describe the procedure in the context of teacher measures; the procedure for the school measures is analogous.

In the first step, we used a dosage-weighted errors-in-variables regression based on equation (2) to obtain unbiased estimates of the pre-test coefficients for each grade. For grades 4 through 8, we used the published reliabilities associated with the 2011 DC CAS.²¹ We then used the measurement-error corrected values of the pre-test coefficients to calculate the adjusted gain for each student in each grade. The adjusted gain is expressed as:

²⁰ Although teachers must teach at least 15 students for DCPS to evaluate them on the basis of individual value added, we included in the regression teachers with 7 to 14 students for two reasons. First, we expected that maintaining more teacher-student links will lead to coefficients on the covariates that are estimated more accurately. Second, we expected that value-added estimates for these teachers will provide useful data to include in the standardization and shrinkage procedures described below. We did not include teachers with fewer than 7 students because estimates for such teachers would be too likely to be outliers, which could skew the standardization and shrinkage procedures. If a teacher had fewer than 7 students in a grade, we reallocated those students to a grade-specific catchall ineligible teacher.

²¹ In the school model, weights were not used when applying errors-in-variables regression because dosage was already accounted for in the school-grade variables. Further, we estimated separate pre-test coefficients for 10th-grade students taking the 8th-grade test in 2010 and those taking the test in 2009, using DC CAS reliabilities from the appropriate year for each type of student. To account for any systematic difference in test score growth between 10th-grade students taking the test three rather than two years previously, the vector of student characteristics in the school regressions includes a binary variable indicating whether a 10th-grade student took the DC CAS in 2009.

$$(3) \quad \hat{G}_{tig} = Y_{tig} - \hat{\lambda}_{2g} Y_{i(g-1)} - \hat{\omega}_{2g} Z_{i(g-1)},$$

and represents the student post-test outcome, net of the estimated contribution attributable to the student's starting position at pre-test.

In the second step, we pooled the data from all grades and used the adjusted gain as the dependent variable in a single equation expressed as:

$$(4) \quad \hat{G}_{tig} = \alpha'_2 \mathbf{X}_{2i} + \eta' \mathbf{T}_{tig} + \varepsilon_{tig},$$

We obtained the grade-specific estimates of teacher effectiveness, $\hat{\eta}$, by applying the WLS regression technique to equation (4).²²

This two-step method will likely underestimate the standard error of $\hat{\eta}$ because the adjusted gain in equation (3) relies on the estimated value of λ_g , which implies that the error term in equation (4) is clustered within grades. This form of clustering typically results in estimated standard errors that are too small because the second-step regression does not account for a common source of variability affecting all students in a grade. In view of the small number of grades, standard techniques of correcting for clustering will not effectively correct the standard errors (Bertrand et al. 2004). Nonetheless, with the large within-grade sample sizes, the pre-test coefficients are likely to be estimated precisely, leading to a negligible difference between the robust and clustering-corrected standard errors.

C. Combining Estimates Across Grades

Both the average and the variability of value-added estimates may differ across grade levels, leading to a potential problem when comparing teachers assigned to different grades or comparing schools with different grade configurations. The main concern is that factors beyond teachers' control—rather than teacher distribution or school effectiveness—may drive cross-grade discrepancies in the distribution of value-added estimates. For example, the standard deviation of adjusted gains might vary across grades as a consequence of differences in the alignment of tests or the retention of knowledge between years. However, we sought to compare all schools or teachers to all others in the regression regardless of any grade-specific factors that might affect the distribution of gains in student performance between years.²³ Because we did not want to penalize or reward teachers simply for teaching in a grade with unusual test properties, we translated grade-level estimates for schools and teachers so that each set of estimates is expressed in a common metric of “generalized” DC CAS points. Below, we describe the procedure in the context of teacher measures; the procedure for school measures is analogous.

We standardized the estimated regression coefficients so that the mean and standard deviation of the distribution of teacher estimates is the same across grades. First, we subtracted from each unadjusted estimate the weighted average of all estimates within the same grade. We then divided

²² In the school model, OLS was applied to a student-level expression comparable to equation (4).

²³ Because each student's entire dosage was accounted for by teachers or schools in a given grade, the information contained in grade indicators would be redundant to the information contained in the teacher or school variables. Therefore, it is not also possible to control directly for grade in the value-added regressions.

the result by the weighted standard deviation within the same grade. To reduce the influence of imprecise estimates obtained from teacher-grade combinations with few students, we based the weights on the number of students taught by each teacher. Finally, we multiplied by the teacher-weighted average of the grade-specific standard deviations, obtaining a common measure of effectiveness on the generalized DC CAS-point scale.

Formally, the value-added estimate expressed in generalized DC CAS points is:

$$(5) \quad \hat{\theta}_{tg} = \frac{\hat{\eta}_{tg} - \bar{\hat{\eta}}_g}{\hat{\sigma}_g} \times \left(\frac{1}{K} \sum_h K_h \hat{\sigma}_h \right),$$

where $\hat{\eta}_{tg}$ is the grade- g estimate for teacher t , $\bar{\hat{\eta}}_g$ is the weighted average estimate for all teachers in grade g , $\hat{\sigma}_g$ is the weighted standard deviation of teacher estimates in grade g , K_h is the number of teachers with students in grade h , and K is the total number of teachers. We excluded the estimates associated with the ineligible teachers (and with the “schools outside DCPS” estimates in the school model). Table III.2 shows the standard deviation of the initial estimates of school and teacher effectiveness derived from the value-added regression as well as the weighted average across all grades produced by Equation 5.

Table III.2. Student-Weighted Standard Deviations of Value-Added Estimates

Model	Grade						Weighted Average
	4	5	6	7	8	10	
Teacher (DCPS and charter schools)							
Math	4.5	4.9	3.6	2.9	2.8	n.a.	4.0
Reading	3.2	3.0	2.4	1.9	2.1	n.a.	2.7
School (DCPS only)							
Math	4.5	4.1	3.5	2.7	2.1	3.5	3.7
Reading	3.2	3.1	2.3	2.0	2.2	2.2	2.7

Notes: Teachers are included in the calculation of the standard deviation for each grade that they teach, weighted by the number of students they teach in that grade.

Aside from putting value-added estimates for teachers onto a common scale, this approach equalizes the distribution of teacher estimates across grades. It does not reflect *a priori* knowledge that the true distribution of teacher effectiveness is similar across grades. Rather, without a way to distinguish cross-grade differences in teacher effectiveness from cross-grade differences in testing conditions, in the test instrument itself, or in student cohorts, we assumed that the distribution of true teacher effectiveness is the same across grades.

To combine effects across grades into a single effect ($\hat{\theta}_t$) for a given teacher, we used a weighted average of the grade-specific estimates (expressed in generalized DC CAS points). We set the weight for grade g equal to the proportion of students of teacher t in grade g , denoted as p_{tg} . We computed the variance of each teacher's estimated effect by using:

$$(6) \quad \text{Var}\left[\hat{\theta}_t\right] = \sum_g p_{tg}^2 \text{Var}\left[\hat{\theta}_{tg}\right],$$

where $\text{Var}\left[\hat{\theta}_{tg}\right]$ is the variance of the grade- g estimate for teacher t . For simplicity, we assumed that the covariance across grades is zero. In addition, we did not account for uncertainty arising because $\hat{\eta}_g$ and $\hat{\sigma}_g$ in equation (6) are estimates of underlying parameters rather than known constants. Both decisions imply that the standard errors obtained from equation (6) will be slightly underestimated. Because combining teacher effects across grades may cause the overall average to be nonzero, we re-centered the estimates on zero before proceeding to the next step.

D. Shrinkage Procedure

To reduce the risk that teachers or schools, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance, we applied the empirical Bayes (EB) shrinkage procedure, as outlined in Morris (1983), separately to the sets of effectiveness estimates for teachers and schools. We frame our discussion of shrinkage in terms of teachers, but the same logic applies to schools. Using the EB procedure, we computed a weighted average of an estimate for the average teacher (based on all students in the model) and the initial estimate based on each teacher's own students. For teachers with relatively imprecise initial estimates based on their own students, the EB method effectively produces an estimate based more on the average teacher. For teachers with more precise initial estimates based on their own students, the EB method puts less weight on the value for the average teacher and more weight on the value obtained from the teacher's own students.

The EB estimate for a teacher is approximately equal to a precision-weighted average of the teacher's initial estimated effect and the overall mean of all estimated teacher effects.²⁴ Following the standardization procedure, the overall mean is zero, with better-than-average teachers having positive scores and worse-than-average teachers having negative scores. We therefore arrived at the following:

$$(7) \quad \hat{\theta}_t^{EB} \approx \left(\frac{\frac{1}{\hat{\sigma}_t^2}}{\frac{1}{\hat{\sigma}_t^2} + \frac{1}{\hat{\sigma}^2}} \right) \hat{\theta}_t,$$

where $\hat{\theta}_t^{EB}$ is the EB estimate for teacher t , $\hat{\theta}_t$ is the initial estimate of effectiveness for teacher t based on the regression model (after combining across grades), $\hat{\sigma}_t$ is the standard error of the estimate of teacher t , and $\hat{\sigma}$ is an estimate of the standard deviation of teacher effects (purged of sampling error), which is constant for all teachers. The term $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2 + 1/\hat{\sigma}^2)]$ must be less than one. Thus, the EB estimate always has a smaller absolute value than the initial estimate—that is,

²⁴ In Morris (1983), the EB estimate does not exactly equal the precision-weighted average of the two values due to a correction for bias. This adjustment increases the weight on the overall mean by $(K - 3)/(K - 1)$, where K is the number of teachers. For ease of exposition, we have omitted this correction from the description given here.

the EB estimate “shrinks” from the initial estimate. The greater the precision of the initial estimate—that is, the larger is $(1/\hat{\sigma}_t^2)$ —the closer $[(1/\hat{\sigma}_t^2)/(1/\hat{\sigma}_t^2 + 1/\hat{\sigma}^2)]$ is to one and the smaller the shrinkage in $\hat{\theta}_t$. Conversely, the smaller the precision of the initial estimate, the greater the shrinkage in $\hat{\theta}_t$. By applying a greater degree of shrinkage to less precisely estimated teacher measures, the procedure reduces the likelihood that the estimate of effectiveness for a teacher falls at either extreme of the distribution by chance. We calculated the standard error for each $\hat{\theta}_t^{EB}$ using the formulas provided by Morris (1983). As a final step, we removed any teachers with fewer than 15 students from the teacher model and any schools with fewer than 100 student tests across both subjects from the school model, and we re-centered the EB estimates on zero.

E. Translating Value-Added Results to Scores for Evaluation Systems

We provided OSSE with the original generalized DC CAS-point score, percentile rankings for individual teachers compared to all DC teachers, and a score converted to a scale from 1.0 to 4.0. OSSE determined the method for converting the score in consultation with the Technical Support Committee, a group of representatives from six DC LEAs. In this system, the average DC teacher (including DCPS and charter school teachers) receives a score of 3.0. The value-added score constitutes half of the total evaluation score for eligible charter school teachers, but each charter LEA determines the exact way in which it will incorporate this information into its evaluation system.

We provided DCPS with value-added results only for DCPS teachers. Because the other components of a teacher’s evaluation in IMPACT are based on DCPS norms, DCPS determined that value-added scores for their teachers should exclude comparisons to charter school teachers. Therefore, we re-centered the scores using only DCPS teachers before we provide value-added scores to DCPS. Consequently, a DCPS teacher with a score of zero generalized DC CAS points is an average teacher relative to other DCPS teachers. We also provided DCPS with percentile rankings compared to DCPS teachers and with a converted score that runs from 1.0 to 4.0 based on a method determined by DCPS. The average DCPS teacher on this scale receives a score of 3.0. The score on the 1.0-to-4.0 scale is incorporated into IMPACT.

Given that the generalized DC CAS-point scores provided to DCPS were shifted to be relative to the average DCPS teacher, DC CAS-point scores provided to DCPS teachers and DC CAS-point scores provided to charter school teachers are not comparable. Likewise, the scores on a scale from 1.0 to 4.0 are not comparable between DCPS teachers and charter school teachers because OSSE and DCPS use different comparison groups and different methods of converting scores.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, vol. 25, no. 1, 2007, pp. 95–135.

Arellano, Manuel. "Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, vol. 49, no. 4, November 1987, pp. 431-34.

Bertrand, M., E. Duflo, and S. Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, vol. 119, no. 1, 2004, pp. 248-275.

Buonaccorsi, John P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman & Hall/CRC, 2010.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." Cambridge, MA: National Bureau of Economic Research, 2011.

CTB/McGraw-Hill. *Technical Report for Spring 2010 Operational Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 2010.

CTB/McGraw-Hill. *Technical Report for Spring 2011 Operational Test Administration of DC CAS*. Monterey, CA: CTB/McGraw-Hill, 2011.

CTB/McGraw-Hill. *Technical Report for the Washington, D.C., Comprehensive Assessment System (DC CAS), Spring 2009*. Monterey, CA: CTB/McGraw-Hill, 2009.

Hanushek, Eric A., John F. Kain, Steven G. Rivkin, and Gregory F. Branch. "Charter School Quality and Parental Decision Making with School Choice." *Journal of Public Economics*, vol. 91, nos. 5-6, 2007, pp. 823–848.

Hock, Heinrich, and Eric Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." Washington, DC: Mathematica Policy Research, June 2012.

Huber, Peter J. "The Behavior of Maximum Likelihood Estimation under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, edited by L.M. LeCam and J. Neyman. Berkeley, CA: University of California Press, 1967.

Isenberg, Eric, and Heinrich Hock. "Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools." Washington, DC: Mathematica Policy Research, 2010.

Isenberg, Eric, and Heinrich Hock. "Design of Value Added Models for IMPACT and TEAM in DC Public Schools, 2010-2011 School Year." Washington, DC: Mathematica Policy Research, 2011.

Kane, Thomas J., and Douglas O. Staiger. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, vol. 16, no. 4, 2002, pp. 91–114.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working paper #14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, Thomas J., and Douglas O. Staiger. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill & Melinda Gates Foundation, 2012.

Lee, P. *Bayesian Statistics: An Introduction*. Second Edition. New York: John Wiley and Sons, 1997.

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, vol. 73, no. 1, April 1986, pp. 13-22.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 67–102.

Meyer, Robert H., "Value-Added Indicators of School Performance: A Primer." *Economics of Education Review*, vol. 16, no. 3, 1997, pp. 283–301.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of American Statistical Association*, vol. 78, no. 381, 1983, pp. 47-55.

Raudenbush, Stephen W. "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 121–129.

Rothstein, Jesse. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, vol. 125, no. 1, 2010, pp. 175–214.

Sanders, William L. "Value-Added Assessment from Student Achievement Data—Opportunities and Hurdles." *Journal of Personnel Evaluation in Education*, vol. 14, no. 4, 2000, pp. 329-339.

Taylor, Eric S., and John H. Tyler. "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers." Working paper #16877. Cambridge, MA: National Bureau of Economic Research, 2011.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, vol. 48, no. 4, 1980, pp. 817-830.

Wooldridge, Jeffrey. *Introductory Econometrics: A Modern Approach*. Fourth Edition. Mason, OH: South-Western/Thomson, 2008.



MATHEMATICA **Policy Research**

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC

Mathematica® is a registered trademark of Mathematica Policy Research